CS 726: Homework #2

Posted: 02/11/2020, due: 02/24/2020 by 5pm on Canvas

Please typeset or write your solutions neatly! If we cannot read it, we cannot grade it.

Note: You can use the results we have proved in class – no need to prove them again.

Q 1. Recall the Gauss-Southwell rule for basic descent methods that we saw in class: $\mathbf{d}_k = -\nabla_{i_k} f(\mathbf{x}_k)\mathbf{e}_{i_k}$, where $i_k = \text{argmax}_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|$ and $\mathbf{e}_{i_k}$ is the vector that has 0 in all coordinates except for i_k, where it equals 1 (it is the ${i_k}^{\text{th}}$ standard basis vector). Same as in the class, we assume that f is L-smooth. Prove that there exists $\alpha > 0$ such that the Gauss-Southwell rule applied for an appropriate step size α_k satisfies:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

How would you choose α_k? What can you say about the convergence of this method (discuss all three cases we have covered in class: nonconvex and bounded below, convex, strongly convex)? [10pts]

Solution. Observe that $\|\mathbf{d}_k\|_2^2 = \max_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|^2 \geq \frac{1}{n}\|\nabla f(\mathbf{x}_k)\|_2^2$. Using smoothness of f, the definition of $\mathbf{d}_k$, and $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$:

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \alpha_k \langle \nabla f(\mathbf{x}_k), \mathbf{d}_k\rangle + \frac{L{\alpha_k}^2}{2}\|\mathbf{d}_k\|_2^2 \\ &= -\alpha_k \max_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|^2 + \frac{L{\alpha_k}^2}{2} \max_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|^2 \\ &= -\left(\alpha_k - \frac{L{\alpha_k}^2}{2}\right) \max_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|^2. \end{aligned}$$

Any choice of α_k such that $\alpha_k - \frac{L{\alpha_k}^2}{2} > 0$ would lead to a descent method. The best choice is the one that maximizes $\alpha_k - \frac{L{\alpha_k}^2}{2}$, which is obtained for $\alpha_k = \frac{1}{L}$. As $\max_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|^2 \geq \frac{1}{n}\|\nabla f(\mathbf{x}_k)\|_2^2$, it leads to the following descent property:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{1}{2nL}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Using the results we have proved in class, we have the following results:

- If f is possibly non-convex and bounded below by $f^* > -\infty$, we have:

$$\min_{0\leq i\leq k} \|\nabla f(\mathbf{x}_i)\|_2 \leq \sqrt{\frac{2nL(f(\mathbf{x}_0) - f^*)}{k+1}}.$$

- If f is convex and $f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x}) > -\infty$, we have:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2nL{R_0}^2}{k+1},$$

  where $R_0 = \max\{\|\mathbf{x} - \mathbf{x}^*\|_2 : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$.

- If f is m-strongly convex and $f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x})$, then:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \left(1 - \frac{m}{nL}\right)^k \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2}.$$

□

Q 2. Exercise 8 from Chapter 3 in Recht-Wright. Notation from the exercise: $\mathbf{x}_\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. You don't need to worry about getting the same constants as stated there, being off by constant factors (up to 4) is fine. [15pts]

Solution.

(a) Yes, $\mathbf{x}_\star^{(\epsilon)}$ is unique, as f_ϵ is strongly convex.

(b) Notice first that $f(\mathbf{z}) \leq f(\mathbf{z}) + \frac{\epsilon}{2R^2}\|\mathbf{z}\|_2^2 = f_\epsilon(\mathbf{z})$. On the other hand:

$$f(\mathbf{x}_\star) = f_\epsilon(\mathbf{x}_\star) - \frac{\epsilon}{2R^2}\|\mathbf{x}_\star\|_2^2 \geq f_\epsilon(\mathbf{x}_\star^{(\epsilon)}) - \frac{\epsilon}{2},$$

as $\mathbf{x}_\star^{(\epsilon)}$ minimizes f_ϵ (and, thus $f_\epsilon(\mathbf{x}_\star) \geq f_\epsilon(\mathbf{x}_\star^{(\epsilon)})$) and $\mathbf{x}_\star$ is contained in a centered ball of radius R (and thus $\|\mathbf{x}_\star\| \leq R$). Combining the upper bound on $f(\mathbf{z})$ and the lower bound on $f(\mathbf{x}_\star)$, the claim follows.

(c) Observe that f_ϵ is smooth with parameter $L + \frac{\epsilon}{R^2}$ and strongly convex with parameter $\frac{\epsilon}{R^2}$. Thus, applying the steepest descent method with the step size $\frac{1}{L+\frac{\epsilon}{R^2}}$ and using the results from the lectures, we have:

$$f_\epsilon(\mathbf{x}_k) - f_\epsilon(\mathbf{x}_\star^{(\epsilon)}) \leq \Big(1 - \frac{\epsilon/R^2}{L+\epsilon/R^2}\Big)^k \frac{(L+\epsilon/R^2)\|\mathbf{x}_0 - \mathbf{x}_\star^{(\epsilon)}\|_2^2}{2} \leq \Big(1 - \frac{\epsilon/R^2}{L+\epsilon/R^2}\Big)^k \frac{(L+\epsilon/R^2)4R^2}{2}.$$

Thus, $f_\epsilon(\mathbf{x}_k) - f_\epsilon(\mathbf{x}_\star^{(\epsilon)}) \leq \frac{\epsilon}{2}$ after at most $2(1 + \frac{LR^2}{\epsilon})\log(4 + \frac{4LR^2}{\epsilon})$ iterations (assuming $\frac{\epsilon/R^2}{L+\epsilon/R^2} \leq 0.7$). Using Part (b), $f(\mathbf{x}_k) - f(\mathbf{x}_\star) \leq \epsilon$ within the same number of iterations.

□

Q 3 (Bregman Divergence)**.** Bregman divergence of a continuously differentiable function $\psi : \mathbb{R}^n \to \mathbb{R}$ is a function of two points defined by

$$D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle .$$

Equivalently, you can view Bregman divergence as the error in the first-order approximation of a function:

$$\psi(\mathbf{x}) = \psi(\mathbf{y}) + \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + D_\psi(\mathbf{x}, \mathbf{y}).$$

(i) What is the Bregman divergence of a simple quadratic function $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$, where $\mathbf{x}_0 \in \mathbb{R}^n$ is a given point? [5pts]

(ii) Given $\mathbf{x}_0 \in \mathbb{R}^n$ and some continuously differentiable $\psi : \mathbb{R}^n - \mathbb{R}$, what is the Bregman divergence of function $\phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{x}_0, \mathbf{x}\rangle$? [5pts]

(iii) Use Part (ii) and the definition of Bregamn divergence to prove the following 3-point identity:

$$(\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n): \quad D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z}\rangle + D_\psi(\mathbf{x}, \mathbf{z}). \qquad [5\text{pts}]$$

(iv) Suppose I give you the following function: $m_k(\mathbf{x}) = \sum_{i=0}^k a_i\psi_i(\mathbf{x})$, where $\psi_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_i\|_2^2 + \langle \mathbf{b}_i, \mathbf{x} - \mathbf{x}_i\rangle$, where $\{a_i\}_{i\geq 0}$ is a sequence of positive reals and $\{\mathbf{b}_i\}_{i=0}^k$, $\{\mathbf{x}_i\}_{i=0}^k$ are fixed vectors from $\mathbb{R}^n$. Define $\mathbf{v}_k = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} m_k(\mathbf{x})$ and $A_k = \sum_{i=0}^k a_i$. Using what you have proved so far, prove the following inequality:

$$(\forall \mathbf{x} \in \mathbb{R}^n): \quad m_{k+1}(\mathbf{x}) \geq m_k(\mathbf{v}_k) + a_{k+1}\psi_{k+1}(\mathbf{x}) + \frac{A_k}{2}\|\mathbf{x} - \mathbf{v}_k\|_2^2. \qquad [5\text{pts}]$$

Solution. Observe that the Bregman divergence of either a linear or a constant function is zero.

(i) Observe that (using the definition of Bregman divergence) for any two functions f, g, we have $D_{f+g}(\mathbf{x}, \mathbf{y}) = D_f(\mathbf{x}, \mathbf{y}) + D_g(\mathbf{x}, \mathbf{y})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Expand the square to get $\frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 = \frac{1}{2}\|\mathbf{x}\|_2^2 - \langle \mathbf{x}, \mathbf{x}_0\rangle + \frac{1}{2}\|\mathbf{x}_0\|_2^2$. Thus, the Bregman divergence of $\frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$ is the same as the Bregman divergence of $\frac{1}{2}\|\mathbf{x}\|_2^2$, which is just:

$$D_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 - \langle \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

(ii) We have already discussed that the Bregman divergence of a linear function is zero and that for any two functions f, g, we have $D_{f+g}(\mathbf{x}, \mathbf{y}) = D_f(\mathbf{x}, \mathbf{y}) + D_g(\mathbf{x}, \mathbf{y}), \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Thus $D_\psi(\mathbf{x}, \mathbf{y}) = D_\phi(\mathbf{x}, \mathbf{y})$.

(iii) Let $\phi(\mathbf{x}) = D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$. As a function of $\mathbf{x}$, all the terms in the definition of $\phi(\mathbf{x})$ apart from $\psi(\mathbf{x})$ are either linear or constant. Thus, for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n : D_\phi(\mathbf{x}, \mathbf{z}) = D_\psi(\mathbf{x}, \mathbf{z})$. By the definition of Bregman divergence (w.r.t. ϕ):

$$\begin{aligned} D_\psi(\mathbf{x}, \mathbf{z}) &= D_\phi(\mathbf{x}, \mathbf{z}) \\ &= D_\psi(\mathbf{x}, \mathbf{y}) - D_\psi(\mathbf{z}, \mathbf{y}) - \langle \nabla_{\mathbf{z}} D_\psi(\mathbf{z}, \mathbf{y}), \mathbf{x} - \mathbf{z}\rangle \\ &= D_\psi(\mathbf{x}, \mathbf{y}) - D_\psi(\mathbf{z}, \mathbf{y}) - \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z}\rangle . \end{aligned}$$

It remains to rearrange the terms in the last equality.

(iv) Consider the Bregman divergence w.r.t. m_k between $\mathbf{x}$ and $\mathbf{v}_k$. Using Part (ii), we have

$$D_{m_k}(\mathbf{x}, \mathbf{v}_k) = D_{\frac{1}{2}\sum_{i=0}^k a_i \|\cdot\|_2^2}(\mathbf{x}, \mathbf{v}_k) = \frac{A_k}{2}\|\mathbf{x} - \mathbf{v}_k\|_2^2.$$

Thus, we have $\forall \mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} m_{k+1}(\mathbf{x}) &= m_k(\mathbf{x}) - m_k(\mathbf{v}_k) + m_k(\mathbf{v}_k) + a_{k+1}\psi_{k+1}(\mathbf{x}) \\ &= \langle \nabla m_k(\mathbf{v}_k), \mathbf{x} - \mathbf{v}_k\rangle + D_{m_k}(\mathbf{x}, \mathbf{v}_k) + m_k(\mathbf{v}_k) + a_{k+1}\psi_{k+1}(\mathbf{x}) \\ &= D_{m_k}(\mathbf{x}, \mathbf{v}_k) + m_k(\mathbf{v}_k) + a_{k+1}\psi_{k+1}(\mathbf{x}), \end{aligned}$$

as $\mathbf{v}_k$ minimizes m_k, and, thus $\nabla m_k(\mathbf{v}_k) = 0$. The rest follows by plugging $D_{m_k}(\mathbf{x}, \mathbf{v}_k) = \frac{A_k}{2}\|\mathbf{x} - \mathbf{v}_k\|_2^2$ into the last equality.

□

Q 4. In class, we have analyzed the following variant of Nesterov's method for L-smooth convex optimization:

$$\begin{aligned} \mathbf{x}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1} \\ \mathbf{v}_k &= \mathbf{v}_{k-1} - a_k\nabla f(\mathbf{x}_k)/L \\ \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k), \end{aligned}$$

where L is the smoothness constant of f, $a_0 = A_0 = 1$, $\frac{{a_k}^2}{A_k} = 1$, $A_k = \sum_{i=0}^k a_i$. We take $\mathbf{x}_0 \in \mathbb{R}^n$ to be an arbitrary initial point and $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0 - \nabla f(\mathbf{x}_0)/L$.

Prove that we can state Nesterov's method in the following equivalent form:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{y}_{k-1} + \frac{a_k}{A_k}\Big(\frac{A_{k-1}}{a_{k-1}} - 1\Big)(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \\ \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k). \end{aligned} \tag{1}$$

Hint: It is helpful to first prove that $\mathbf{y}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k$. [10pts]

Solution. Let us first prove the statement from the hint. Using the definitions of $\mathbf{x}_k$ and $\mathbf{v}_k$:

$$\begin{aligned} \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k) \\ &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\Big(\mathbf{v}_{k-1} - \frac{A_k}{a_k}\frac{1}{L}\nabla f(\mathbf{x}_k)\Big) \\ &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k, \end{aligned}$$

where the last line is by $\frac{{a_k}^2}{A_k} = 1$ and $\mathbf{v}_k = \mathbf{v}_{k-1} - a_k \nabla f(\mathbf{x}_k)/L$.

It follows that we can express $\mathbf{v}_{k-1}$ as $\mathbf{v}_{k-1} = \frac{A_{k-1}}{a_{k-1}}\Big(\mathbf{y}_{k-1} - \frac{A_{k-2}}{A_{k-1}}\mathbf{y}_{k-2}\Big)$. Using the definition of $\mathbf{x}_k$:

$$\begin{aligned}
\mathbf{x}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\frac{A_{k-1}}{a_{k-1}}\Big(\mathbf{y}_{k-1} - \frac{A_{k-2}}{A_{k-1}}\mathbf{y}_{k-2}\Big)\\
&= \mathbf{y}_{k-1} + \frac{a_k}{A_k}\Big(\Big(\frac{A_{k-1}}{a_{k-1}} - 1\Big)\mathbf{y}_{k-1} - \frac{A_{k-2}}{A_{k-1}}\mathbf{y}_{k-2}\Big)\\
&= \mathbf{y}_{k-1} + \frac{a_k}{A_k}\Big(\frac{A_{k-1}}{a_{k-1}} - 1\Big)(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}),
\end{aligned}$$

as $\frac{A_{k-1}}{a_{k-1}} - 1 = \frac{A_{k-1} - a_{k-1}}{a_{k-1}} = \frac{A_{k-2}}{A_{k-1}}$. □

Q 5 (Coding Assignment)**.** In the coding assignment, we will compare different optimization methods discussed in class on the following problem instance: $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$, where $f(\mathbf{x}) = \frac{1}{2}\langle \mathbf{M}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$, $\mathbf{b}$ is a vector whose first coordinate is $1 - \frac{1}{n}$ while the remaining coordinates are $\frac{1}{n}$, and $\mathbf{M}$ is the same matrix we saw in Q 8 of Homework #1. We will take the dimension to be $n = 200$. Matrix $\mathbf{M}$ and vector $\mathbf{b}$ can be generated using the following Matlab code:

```
k = n;
M = diag(2*[ones(k, 1); zeros(n-k, 1)], 0)...
    + diag([-ones(k-1, 1); zeros(n-k, 1)], -1)...
    + diag([-ones(k-1, 1); zeros(n-k, 1)], 1);
M(n,1) = - 1;
M(1,n) = -1;
b = -1/n * ones(n, 1);
b(1) = b(1) + 1;
```

Observe that you can compute the minimizer $\mathbf{x}^*$ of f given $\mathbf{M}$ and $\mathbf{b}$, and thus you can also compute $f(\mathbf{x}^*)$. It is possible to show that the top eigenvalue of $\mathbf{M}$ is $L = 4$.

Implement the following algorithms:

1. Steepest descent with the constant step size $\alpha_k = 1/L$.
2. Steepest descent with the exact line search.
3. Lagged steepest descent, defined as follows: Let α_k be the exact line search steepest descent step size corresponding to the point $\mathbf{x}_k$. Lagged steepest descent updates the iterates as: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_{k-1}\nabla f(\mathbf{x}_k)$ (i.e., the step size "lags" by one iteration).
4. Nesterov's method for smooth convex minimization.

Initialize all algorithms at $\mathbf{x}_0 = \mathbf{0}$. All your plots should be showing the optimality gap $f(\mathbf{x}) - f(\mathbf{x}^*)$ (with $\mathbf{x} = \mathbf{y}_k$ for Nesterov and $\mathbf{x} = \mathbf{x}_k$ for all other methods) on the y-axis and the iteration count on the x-axis. The y-axis should be shown on a logarithmic scale (use `set(gca, 'YScale', 'log')` after the figure command in Matlab).

(i) Use a single plot to compare steepest descent with constant step size, steepest descent with the exact line search, and Nesterov's algorithm. Use different colors for different algorithms and show a legend with descriptive labels (e.g., SD:constant, SD:exact, and Nesterov). Discuss the results. Do you see what you expect from the analysis we saw in class?

(ii) Use a single plot to compare Nesterov's algorithm to lagged steepest descent. You should, again, use different colors and a legend. What can you say about lagged steepest descent? How does it compare to Nesterov's algorithm?

(iii) Modify the output of Nesterov's algorithm and lagged steepest descent: you should still run the same algorithms, but now your plot at each iteration k should show the lowest function value up to iteration k for each of the two algorithms. Discuss the results.

You should turn in both the code (as a text file) and a pdf with the figures produced by your code together with the appropriate answers to the above questions. [45pts]

Note: Your code needs to compile without any errors to receive any points for the coding assignment.

Solution. Here, we only provide the plots and a brief discussion, for reference.

(i) The output of the code is provided in Fig. 1. We can observe that, in agreement with the theoretical results, Nesterov's AGD is faster than either variant of SD. SD with the exact line search is only slightly faster than SD with a constant step size. Unlike either variant of SD, Nesterov's AGD is not a descent method – it exhibits "ripples" that are a consequence of more aggressive steps and "overshooting."

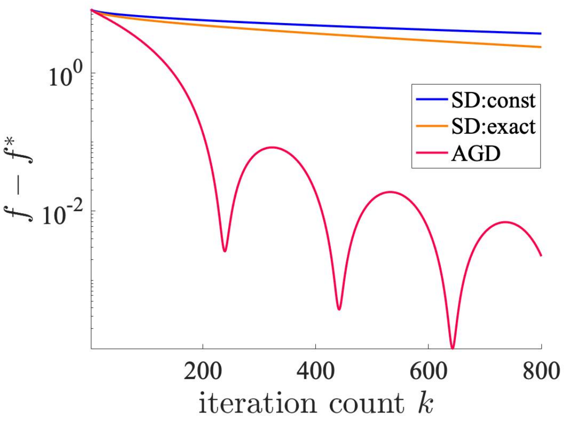

Figure 1: Comparison of steepest descent with constant step size $1/L$, steepest descent with the exact line search, and Nesterov's AGD.

(ii) The output of the code is provided in Fig. 2. LSD is a chaotic method: not only is it not a descent method, but unlike for AGD, even its upper envelope is not a decreasing function of the iteration count. After sufficiently many iterations, however, LSD finds points with much lower function values and it converges faster.

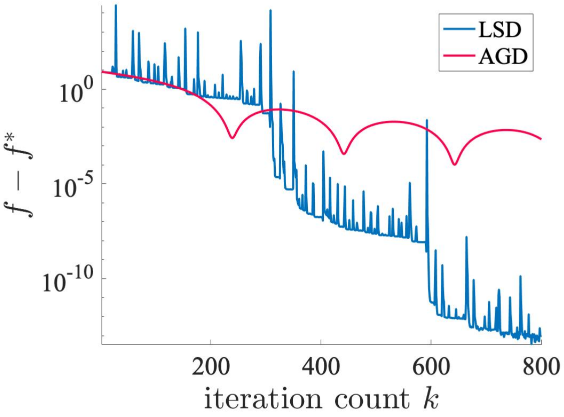

Figure 2: Comparison of lagged steepest descent with Nesterov's AGD.

(iii) The output of the code is provided in Fig. 2. It is easier to compare LSD to AGD when one considers the points with the lowest function value seen so far. Note that in practice, whenever the function can be evaluated, we can always opt to output the best point an algorithm constructs up to a given iteration. In the initial $\sim$ 300 iterations, LSD is either worse or marginally better than AGD. We expect such a result, due to the lower bound we have proved in class (and optimality of AGD), which in this case would apply for the first constant fraction of n iterations. However, once the iteration count is sufficiently higher than n, LSD makes much faster progress towards the minimum function value than AGD.

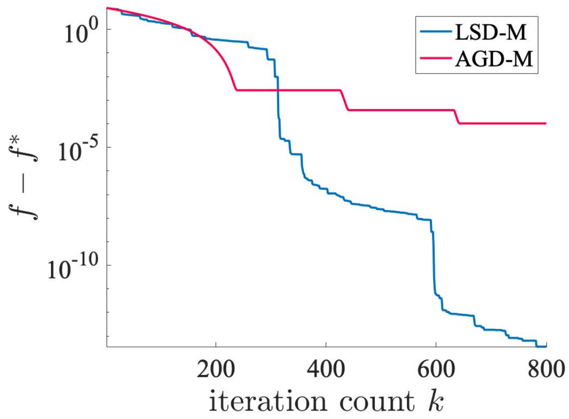

Figure 3: Comparison of lagged steepest descent with Nesterov's AGD, observing the best points up to running iterations.

□